Journal of Applied Ecology 2006
43, 802–815

The ghost of past species occurrence: improving species distribution models for presence-only data

MICHAEL LÜTOLF,*† FELIX KIENAST* and ANTOINE GUISAN†

**Swiss Federal Research Institute WSL, Zürcherstrasse 111, CH-8903 Birmensdorf, Switzerland; and †Laboratory for Conservation Biology (LBC), Department of Ecology and Evolution, University of Lausanne, BB, CH-1015 Lausanne, Switzerland*

Summary

1. Model-based approaches have been used increasingly in conservation biology over recent years. Species presence data used for predictive species distribution modelling are abundant in natural history collections, whereas reliable absence data are sparse, most notably for vagrant species such as butterflies and snakes. As predictive methods such as generalized linear models (GLM) require absence data, various strategies have been proposed to select pseudo-absence data. However, only a few studies exist that compare different approaches to generating these pseudo-absence data.

2. Natural history collection data are usually available for long periods of time (decades or even centuries), thus allowing historical considerations. However, this historical dimension has rarely been assessed in studies of species distribution, although there is great potential for understanding current patterns, i.e. the past is the key to the present.

3. We used GLM to model the distributions of three 'target' butterfly species, *Melitaea didyma*, *Coenonympha tullia* and *Maculinea teleius*, in Switzerland. We developed and compared four strategies for defining pools of pseudo-absence data and applied them to natural history collection data from the last 10, 30 and 100 years. Pools included: (i) sites without target species records; (ii) sites where butterfly species other than the target species were present; (iii) sites without butterfly species but with habitat characteristics similar to those required by the target species; and (iv) a combination of the second and third strategies. Models were evaluated and compared by the total deviance explained, the maximized Kappa and the area under the curve (AUC).

4. Among the four strategies, model performance was best for strategy 3. Contrary to expectations, strategy 2 resulted in even lower model performance compared with models with pseudo-absence data simulated totally at random (strategy 1).

5. Independent of the strategy model, performance was enhanced when sites with historical species presence data were not considered as pseudo-absence data. Therefore, the combination of strategy 3 with species records from the last 100 years achieved the highest model performance.

6. *Synthesis and applications.* The protection of suitable habitat for species survival or reintroduction in rapidly changing landscapes is a high priority among conservationists. Model-based approaches offer planning authorities the possibility of delimiting priority areas for species detection or habitat protection. The performance of these models can be enhanced by fitting them with pseudo-absence data relying on large archives of natural history collection species presence data rather than using randomly sampled pseudo-absence data.

Key-words: generalized linear model, historical species data, model performance, predictive species distribution model

Journal of Applied Ecology (2006) **43**, 802–815
doi: 10.1111/j.1365-2664.2006.01191.x

Correspondence: Felix Kienast, Swiss Federal Research Institute WSL, Zürcherstrasse 111, CH-8903 Birmensdorf, Switzerland (fax +41 1 739 22 15).

Introduction

Predictive species distribution models make use of statistical techniques and geographical information technology to simulate the spatial distribution of species (reviewed by Guisan & Zimmermann 2000). Over the last few decades, predictive modelling has become a prominent tool with which to assess the impacts of climate change (Busby 1991; Guisan & Theurillat 2000; Dirnböck, Dullinger & Grabherr 2003) and land-use change (Pearson, Turner & Drake 1999; Osborne, Alonso & Bryant 2001; Dirnböck, Dullinger & Grabherr 2003) on species distributions. Moreover, it has been applied to detect new populations of rare and endangered species in the field (Engler, Guisan & Rechsteiner 2004) and has served to identify areas with a high potential for (re)colonization (Pearce & Lindenmayer 1998; Hirzel *et al*. 2002). Hence, species distribution models have become important tools for nature conservation planning (Peterson *et al*. 2000; Guisan & Thuiller 2005; Whittaker *et al*. 2005).

Generalized linear modelling (GLM; McCullagh & Nelder 1989) is a common way to model species' distributions based on presence–absence data. Although the presence of a species can be unambiguously confirmed in the field, provided that its identification is correct, the absence of a species is much harder to testify. Species absence can result from: (i) a species being undetected at the visited site (MacKenzie *et al*. 2003); (ii) the species being only temporary absent from the site (for example not yet recolonized after an extinction event and a biannual cycle); (iii) the site not yet being colonized (invasive species; Hirzel, Helfer & Metral 2001); (iv) the environment being unsuitable. For habitat modelling purposes only the latter case represents a 'real' absence. Misclassification of a site as an absence may be more frequent for small populations, for example with rare and endangered species (McArdle 1990; Kéry 2002). Kéry (2002) calculated for a snake species that up to 34 successive unsuccessful visits of a location have to be undertaken before it can be assumed with 95% confidence that the species is absent.

'Presence-only' data are very common sources of species distribution information, present in the form of natural history collections (NHC; Graham *et al*. 2004) such as museum collections, herbaria, floras and assembled in reports of field trips. Such data are often available for quite long time periods (e.g. up to 100 years), allowing changes in species presence to be tracked. Indeed, nearly all historical data available originate from NHC databases, and they constitute a unique source for historical analyses. As historical records mostly originate from opportunistic sampling, they show sampling biases (Graham *et al*. 2004). For instance, rare species are often more abundant in collections than common species, and certain localities have been more frequently visited than others. Further difficulties when working with historical data might arise when no information is available on the locality and/or date on which the species was found. Moreover, the spatial location of the find might only be given approximately, or there might be problems with the nomenclature resulting in ambiguous use of synonyms (Graham *et al*. 2004). Furthermore, there is no way to verify the data and no knowledge usually exists about data quality. However, even though NHC data are usually sampled without design, and thus are often thought of as being useless for robust statistical analysis and modelling, they are numerous and thus provide an amazing source of species distribution information.

There are many ways to model species distribution with presence-only data. One category of modelling techniques is based on Hutchinson's (1957) concept of the ecological niche and uses presence-only data (BIOCLIM, Busby 1991; ANUCLIM, Houlder *et al*. 2000; PCA-based technique, Robertson, Caithness & Villet 2001; Ecological niche factor analysis (ENFA), Hirzel *et al*. 2002). A second category solves the problem of missing absence data by randomly sampling cells in the study area or by using more sophisticated procedures to select such 'pseudo-absence data' (Zaniewski, Lehmann & Overton 2002; Engler, Guisan & Rechsteiner 2004). Random selection has not often resulted in the best models. Hirzel *et al*. (2002) mention that the inclusion of doubtful absence data into a GLM may cause a prominent decrease in model performance. Hence a well considered choice of pseudo-absence data has to be made prior to model fitting.

Even though NHC data are very abundant, we are not aware of any study that has taken advantage of this source of information for predicting species distribution over a broader geographical extent, where species have mostly not been sampled systematically. Because we apply static distribution models, we assume a relative equilibrium between the environment (e.g. land use and climate) and the observed species pattern that is only valid for a limited period of time (Guisan & Theurillat 2000; Guisan & Zimmermann 2000). Hence historical species presence data should not be used for predicting recent species' distributions. However, we highlight their use as a 'ghost of past presence data' in order to support the selection of pseudo-absence data.

In this study we assessed two main issues. First, we tested various strategies of resampling NHC data, as a way of generating pseudo-absence data to be used in modelling analyses. Secondly, we tested the importance of historical data (past presence data) and, in particular the best way to incorporate them into the modelling process.

We tested four strategies for generating initial pools of pseudo-absence data, using three butterfly species, *Melitaea didyma*, *Coenonympha tullia* and *Maculinea teleius*, as models. The strategies made potential use of the spatial and temporal presence information of another 197 butterfly species in the study area. Initial pools for a given target species were composed of sites where the modelled species had not been recorded (S1), sites where butterfly species other than the modelled species were present (S2), sites that did not contain butterfly

species but had habitat requirements similar to the modelled species (S3), sites generated by combining the second and third strategies (S4). The strategies were applied to NHC data sets containing species records from either 10 (1991–2000), 30 (1971–2000) or 100 (1901–2000) years to obtain pseudo-absence data. Using the data sets of 30 or 100 years, species presence data from before 1991 were not used as presence data in the binomial GLM. However, such 'ghost' presence data were excluded as possible pseudo-absence data by the strategies used. Therefore the design established for this study not only allowed the influence of different selection strategies to be assessed but also the impact of using different large historical species data sets on model performance.

Materials and methods

STUDY AREA

The study was conducted in Switzerland (41 293 km^2). The northern part of the country (the Jura Mountains, Plateau and northern foothills of the Alps) is dominated by maritime climatic conditions, whereas the south of the Alps is influenced by Mediterranean conditions. High annual rainfall occurs in the Jura Mountains and across the Alps. However, inner-alpine valleys orientated west–east have low precipitation values because they are situated in the rain shadow. Mean summer temperatures on the Plateau are between 19 °C and 21 °C.

Politically, Switzerland is divided into 26 cantons that are further split into the smallest political entities, the communes. Data used for analyses in the present study were derived from and refer to the 2836 communes present in 2003.

SPECIES DISTRIBUTION DATA AND ECOLOGY

The data comprised records of 200 butterfly species (Rhopalocera) originating from observations reported by amateur and professional entomologists between 1901 and 2000 (used for model calibration) and between 2001 and 2005 (used for model evaluation). Species data were spatially aggregated at the communal level (political entities) because (i) most of the environmental predictors were recorded at this level and (ii) historical field observations had implied uncertainties concerning their spatial accuracy that could be levelled out by aggregation at the communal level (i.e. more data were available at this resolution). Three sets of species data were created that differed in the length of the time periods from which the presence data of the 200 species originated. The first set covered species presence data from 1991 to 2000 (10 years), the second from 1971 to 2000 (30 years) and the third from 1901 to 2000 (100 years). The longer the time period the more communes there were with species presence data in the data sets.

TARGET SPECIES

From the 200 available species, three species, *Maculinea teleius*, *Coenonympha tullia* and *Melitaea didyma* (Table 1), were selected as model species for the analyses and are referred to as target species. Both *M. teleius* and *C. tullia* are ecologically strongly related with fenlands, whereas

Table 1. List of the three target species modelled by GLM and the auxiliary species used in strategies 3 and 4 to define the initial pools of pseudo-absence data. The distribution of *M. teleius* was modelled with *C. tullia* as auxiliary species and vice versa. The current status of vulnerability is given according to the International Union for Conservation of Nature and Natural Resources (IUCN) and the Swiss Red List

Scientific name*	Common name	IUCN‡	Red List§
Target species			
Maculinea teleius Bergsträsser 1779	Scarce large blue	LR/nt	2
Coenonympha tullia Müller 1764	Great heath	–	2
Melitaea didyma Esper 1779	Spotted fritillary	–	3
Auxiliary species for *M. teleius* and *C. tullia* (not including either *C. tullia* or *M. teleius*)			
Maculinea alcon Denis & Schiffermüller 1775	Alcon large blue	LR/nt	1
Maculinea nausithous Bergsträsser 1779	Dusky large blue	LR/nt	2
Coenonympha hero Linnaeus 1761	Scarce heath	–	1†
Coenonympha oedippus Fabricius 1787	False ringlet	LR/nt	1†
Boloria aquilonaris Stichel 1908	Cranberry fritillary	–	2
Colias palaeno europome Esper 1777	Moorland clouded yellow	–	3
Vacciniina optilete Knoch 1781	Cranberry blue	–	–
Auxiliary species for *M. didyma*			
Aricia agestis Denis & Schiffermüller 1775	Brown argus	–	3
Aricia artaxerxes Fabricius 1793	Northern brown argus	–	–
Lycaeides idas Linnaeus 1761	Idas blue	–	3
Melitaea cinxia Linnaeus 1758	Glanville fritillary	–	2
Pseudophilotes baton Bergsträsser 1779	Baton blue	–	3

*Taxonomy according to Ebert & Rennwald (1993).
‡http://www.iucnredlist.org (accessed 15 May 2006); LR/nt, lower risk/near threatened.
§Described in Gonseth (1994). 1, risk of extinction; 2, heavily vulnerable; 3, vulnerable; †extinct.

M. didyma is mainly a dry grassland species. All three species appear on the Swiss red lists as vulnerable or endangered. This status originates from past and ongoing land-use changes that have led to a loss of suitable habitats. Such changes include drainage, fertilization, extension of pasture and afforestation (Gonseth 1987; Ebert & Rennwald 1993).

AUXILIARY SPECIES

Two selection strategies for pseudo-absence data presented below made use of the spatial distribution data of species that show habitat requirements similar to the target species. These species are referred to as auxiliary species (Table 1). For *M. teleius*, we selected *C. tullia* and seven species appearing dominantly in wet habitats as auxiliary species. In the case of *C. tullia*, *M. teleius* was selected together with the same seven species. The two species *Coenonympha hero* and *C. oedippus* have not been observed since the 1980s and are therefore considered extinct in Switzerland. Auxiliary species for *M. didyma* included five species selected in the ecofaunal database of Walter & Schneider (2003) because of identical habitat characteristics (dry grassland, scree) as well as similar altitudinal range and threats, both described in Gonseth (1987).

ENVIRONMENTAL PREDICTORS

For every commune, we prepared predictors related to (i) climate, (ii) land use, (iii) agricultural structures, (iv) communal structures and (v) spatial location of the communes (Table 2). The latter was included to account for a spatial trend in the data. Predictors of land use, agricultural and communal structures represented either the state of the landscape (e.g. wetland and intensively cultivated areas) or a driving force affecting the landscape (e.g. number of tractors and classification as an agglomeration). None of the predictors exhibited a correlation greater than 0·6.

CLIMATE

Climatic predictors included mean July temperature and mean July water budget (precipitation – potential evapotranspiration). These were available as geographic information system (GIS) grid layers with a resolution of 25×25 m. They were initially derived from a digital elevation model (DEM; Swisstopo 2005) and meteorological data from the period 1961–90 (for statistical methods see Zimmermann & Kienast 1999). Means of the climatic predictors were calculated for every commune using ArcGIS (ESRI Inc., Redlands, CA).

DEMOGRAPHY AND LAND USE

Population and agricultural censuses for every Swiss commune were carried out in 2000. Population data were obtained from Schuler, Ullmann & Haug (2002). Census data on agriculture, inhabited apartments and employment were obtained from the Swiss Federal Statistical Office. As the communes differed in size, we related the data to either the communal or agricultural area. The resulting percentages then allowed comparisons between the communes.

STATISTICAL MODELLING

We fitted binomial GLM with a logistic link function in the R statistical software (R1·9·1 A Language &

Table 2. Description of the environmental predictors. Type refers to the numbering used in the main text. The spatial reference indicates if the variable was expressed as a proportion of the agricultural (AA) or communal area (CA)

Type	Description	Spatial reference	Unit
1	Mean July temperature	–	°C
1	Mean July water budget (precipitation – potential evapotranspiration)	–	mm
2	Intensively cultivated area (including arable land, meadows that are ploughed from time to time, vineyards and orchards)	CA	%
2	Area of natural (native) meadows	CA	%
2	Area of wetlands	CA	%
3	Number of farms with an agricultural area of 0–5 ha	CA	n ha^{-1}
3	Number of farms with an agricultural area of 5–10 ha	CA	n ha^{-1}
3	Number of farms with an agricultural area of 10–20	CA	n ha^{-1}
3	Number of farms with an agricultural area of > 20 ha	CA	n ha^{-1}
3	Number of full-time farmers	AA	n ha^{-1}
3	Number of part-time farmers	AA	n ha^{-1}
3	Number of farmers between 35 and 50 years	AA	n ha^{-1}
3	Number of farmers between 50 and 65 years	AA	n ha^{-1}
3	Number of livestock	AA	n ha^{-1}
3	Number of tractors	AA	n ha^{-1}
4	Number of inhabited apartments	CA	n ha^{-1}
4	Commune classified as an agglomeration* or not	–	–
5	x coordinate of the centroid of the commune	–	–
5	y coordinate of the centroid of the commune	–	–

*Definition according to Schuler (1997).

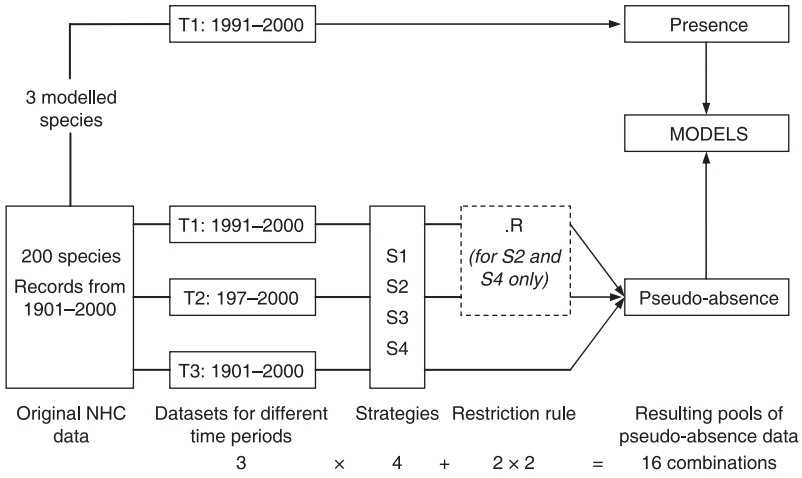

Fig. 1. Analysed framework showing the different species data sets and the different resampling strategies used to derive the pseudo-absence data.

Environment, © 2004). The predictors remaining in the final models were selected using the R-function STEP-AIC, a stepwise backward selection procedure based on the Akaike information criterion (AIC).

For every model, communes where the target species were recorded in the time period 1991–2000 were used as presence data. As static modelling assumes equilibrium between the environment and observed species patterns (Guisan & Zimmermann 2000), we did not include presence data from before 1991 to fit with the environmental predictors recorded in 2000.

STRATEGIES FOR SELECTING PSEUDO-ABSENCE DATA

Binary GLM not only require species presence data but also absence data. However, such absence data are rarely collected because attention is usually concentrated on finding new populations or reconfirming existing records. As we lacked validated absence data, pools of initial pseudo-absence communes were delineated, out of which communes were then randomly selected as absence data of the target species. Initial pools of pseudo-absence data were obtained by applying four different selection strategies (denoted as S1, S2, S3 and S4) to each data set.

Strategy 1 (S1)

From the total of communes present in the area of investigation, only those for which no presence data of the target species were reported remained in the pool.

Strategy 2 (S2)

As the data used did not originate from exhaustive inventories, not all communes showed species records. In strategy 2 only communes for which presence data of species other than the target species were mentioned remained in the pool of pseudo-absence data.

Strategy 3 (S3)

In the pool of pseudo-absence data of strategy 3, only communes where no target species records nor auxiliary species records were present remained.

Strategy 4 (S4)

Strategy 4 was a combination of strategy 3 and strategy 2. Therefore, communes that contained presence data of neither the target nor the auxiliary species but showed at least one species record remained in the pool of pseudo-absence data.

The four selection strategies were applied to each of the three butterfly species data sets comprising the distribution data of all 200 butterfly species over varying time periods (Fig. 1 and Table 3). Time period 1 (T1) contained butterfly species records from 1991 to 2000, which was identical to the time period from which the presence data of the target species originated. Time period 2 (T2) covered 30 years (1971–2000) and time period 3 (T3) 100 years (1901–2000). Hence, the combination of the four strategies with the three time periods not only allowed comparisons of the performance of different selection strategies, but also enabled us to assess the performance of different large historical species data sets.

By applying strategies 1 and 3 to the time periods, the initial pools of pseudo-absence data contained fewer communes the longer the time period lasted, as more presence data of the target species (in strategy 1) or target and auxiliary species (in strategy 3) were available (Table 3). In comparison, the number of communes increased in the initial pools of strategies 2 and 4, as more communes with at least one species record were retained in the pools the longer the period (Table 3). Therefore, by increasing the time period from 10 to 30 and 100 years, we assessed the impact of accounting for former presence data on model performance in strategies 1 and 3. On the other hand, in strategies 2 and 4 we

Journal of Applied Ecology, **43**, 802–815

Table 3. Summary list of all 16 model types, for which the number of communes remaining in the initial pools of pseudo-absence communes (Pool) is indicated. Additional, for strategies 2 and 4 the number of communes without any species records is given (No rec.)

		M. didyma		*C. tullia*		*M. teleius*	
Model type	Criteria communes have to meet to be selected as pseudo-absence	Pool	No rec.	Pool	No rec.	Pool	No rec.
S1–T1	No presence of the target species reported for 1991–2000	2744	0	2820	0	2810	0
S1–T2	No presence of the target species reported for 1971–2000	2685	0	2783	0	2792	0
S1–T3	No presence of the target species reported for 1901–2000	2531	0	2729	0	2752	0
S2–T1	Presence of butterfly species other than the target species reported for 1991–2000	1056	1688	1132	1688	1122	1688
S2–T2	Presence of butterfly species other than the target species reported for 1971–2000	1492	1193	1590	1193	1599	1193
S2–T3	Presence of butterfly species other than the target species reported for 1901–2000	1582	949	1780	949	1803	949
S3–T1	No target and no auxiliary species records reported for 1991–2000	2571	0	2648	0	2648	0
S3–T2	No target and no auxiliary species records reported for 1971–2000	2360	0	2507	0	2507	0
S3–T3	No target and no auxiliary species records reported for 1901–2000	2163	0	2374	0	2374	0
S4–T1	No target and no auxiliary species records but at least one butterfly record for 1991–2000	883	1688	960	1688	960	1688
S4–T2	No target and no auxiliary species records but at least one butterfly record for 1971–2000	1167	1139	1314	1139	1314	1139
S4–T3	No target and no auxiliary species records but at least one butterfly record for 1901–2000	1214	949	1425	949	1425	949
S2–T1.R	Same as in S2–T1, additionally no target species records reported for 1901–90	904	1688	1064	1688	1089	1688
S2–T2.R	Same as in S2–T2, additionally no target species records reported for 1901–70	1352	1193	1540	1193	1566	1193
S4–T1.R	Same as in S4–T1, additionally no target species records reported for 1901–90	650	1688	783	1688	783	1688
S4–T2.R	Same as in S4–T2, additionally no target species records reported for 1901–70	1006	1193	1196	1193	1196	1193

primarily assessed the impact of not taking into account communes without any species records. To account for former presence data in these strategies, four additional model types were established (Fig. 1). Therefore, we restricted the initial pools of the model types S2–T1 and S2–T2 so that no presence data of the target species originating from the period 1901–90 remained (new model types S2–T1.R and S2–T2.R). Similarly, in the initial pools of the model types S4–T1 and S4–T2, communes containing records of target or auxiliary species from the period 1901–90 were not retained as pseudo-absence data any more (S4–T1.R and S4–T2.R). Hence, for every target species 16 model types were performed (Table 3).

There is currently an ongoing discussion about how severe absence data that lie outside a species' known distribution (i.e. 'naughty noughts'; Thuiller *et al*. 2004; Maggini *et al*. 2006) influence species' response curves and thus spatial predictions. However, as in our case the species were widely spread in a relatively small study area we did not include further strategies to limit pseudo-absence communes based on rules for removing naughty noughts.

HYPOTHESIZED MODEL PERFORMANCES

As other studies have shown, predictive distribution models based on a random selection of pseudo-absence data could often be improved using other modelling approaches (Engler, Guisan & Rechsteiner 2004). Hence we expected strategy 1 to show the lowest performance. We further hypothesized that model performance should increase from strategy 1 to strategy 2, as we would expect the target species to be more likely to be absent at the same time as a lot of other species were reported. Even though this assumption was expected to hold better for communes where many species as well as the target species had been found, we did not investigate such a differentiation further and defined the minimum number of species presence data in the communes as one. Furthermore, we hypothesized that the performance would also increase from strategy 1 to strategy 3, where presence data of auxiliary species were not considered as pseudo-absence data. As the modelled species and the auxiliary species share most habitat characteristics, we expected communes where auxiliary species were recorded to provide possible presence data for the target species. In such communes, the target species might have been temporally absent, not been detected or not yet colonized the site. Therefore, presence data of auxiliary species indicated possible recent or future habitats of the target species. As strategy 4 was a combination of strategy 2 and strategy 3, we assumed that it should outperform strategies 2 and 3.

Concerning the different time periods used, we hypothesized that the longer the time period the better the model performance. By considering longer time periods, fewer communes where target or auxiliary species had occurred remained in the pools of initial pseudo-absence data. Even though landscape changes had taken place, it is possible that a site where a target or auxiliary species had once been recorded still offered a suitable habitat. In Switzerland we would expect less change to have occurred in the Alps and therefore that former presence data would be more likely to be reconfirmed in this area than in the Swiss Midlands. Therefore, when longer time periods were considered, communes where species could potentially be reconfirmed were excluded from the pools of pseudo-absence data. Hence we expected a better discrimination in the logistic regression and thus a better model performance.

Whereas the pools of pseudo-absence data in strategies 2 and 4 still contained former presence data of target and auxiliary species, these were removed in the restricted pools. Therefore we assumed that models based on the restricted pools of pseudo-absence data would outperform their corresponding models that relied on the non-restricted pools.

MODEL EVALUATION

Models were evaluated on the basis of the training data set by using resampling techniques (Guisan & Zimmermann 2000), as at the communal level no reliable presence–absence data were available for butterfly species in Switzerland. We calculated the adjusted D^2, the maximum Kappa and the area under the curve (AUC) as measures of model performance, resulting in species-specific hierarchies of model types showing increasing model performance. The robustness of the resulting rankings was assessed further using recent species presence data.

We fitted 1000 GLM for each model type to account for the variance resulting from sampling pseudo-absence data at random. The fit of every model run was characterized by the percentage of deviance reduction (equivalent to the variance reduction in least-squares models) explained by the GLM. We calculated the adjusted D^2 (Guisan & Zimmermann 2000), which takes into account the number of observations and parameters used to build the model. Each model run was evaluated using a leave-one-out jack-knife procedure (Manly 1997; Guisan & Zimmermann 2000; Jaberg & Guisan 2001). Therefore, using the same set of predictors selected in the final model, new GLM were fitted on data sets reduced by a single observation at a time. This procedure was repeated until every observation of the source data set was left out once. At each run, the fitted model was used to predict the response for the excluded observation. Predictions were reclassified to presence–absence (1–0) for all threshold values between 0·05 and 0·95 by increments of 0·05, and confusion matrices with the observed presence–absence information were generated (Fielding & Bell 1997). The Kappa statistic (Cohen 1960; Fielding & Bell 1997) was calculated for every confusion matrix, and the maximum Kappa value (max. Kappa) was assigned to the model (Guisan & Hofer 2003; Engler, Guisan &

Rechsteiner 2004). Moreover, we used the threshold-independent receiver operating characteristic (ROC) method (Fielding & Bell 1997) to derive the AUC value as a measure of prediction success. The AUC takes values between 0·5 and 1·0, where a value of 0·5 indicates a chance performance and a value of 1·0 represents a model that perfectly separates presence and absence data. We used a one-tailed Mann–Whitney *U*-test to test pairs of strategies and/or time periods for increases in model performance (measured by the adjusted D^2, the max. Kappa and the AUC) as stated in our hypotheses. Hence for each performance measure rankings of the 16 model types resulted.

To assess further the robustness of these rankings, obtained by internal evaluation (resampling techniques), we additionally extracted the probabilities of occurrence for independent species presence data that (i) were recorded between 2001 and 2005 and (ii) did not appear in communes with target species presence data used for model calibration. We averaged the occurrence probabilities resulting from the 1000 fitted models for every model type and ranked the types in ascending order. For each species we defined the agreement between the different rankings (defined by the adjusted D^2, max. Kappa, AUC and occurrence probabilities) of the model types using Spearman rank correlation. Additionally, we compared the rankings of each performance measure and the occurrence probability across the species. A Spearman correlation of 1 represents an identical order of the model types, whereas values close to 0 indicate no agreement.

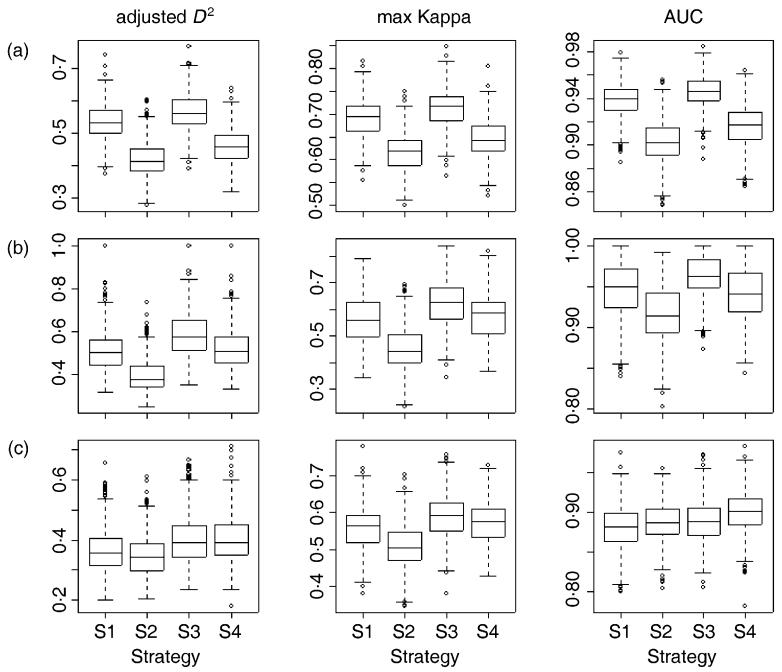

Fig. 2. Model performances for the four strategies S1, S2, S3 and S4 for (a) *Melitaea didyma*, (b) *Coenonympha tullia* and (c) *Maculinea teleius.* Model performance is expressed by the adjusted D^2, max. Kappa and AUC for the time period T1 (1991–2000). Each box-plot represents the results of 1000 model runs with randomly selected pseudo-absence communes from the respective initial pool.

Results

MODEL PERFORMANCE OF THE SELECTION STRATEGIES

The differences in model performance between strategies taking into account communes without species records (strategies 1 and 3) and strategies where such communes did not remain in the initial pool of pseudo-absence data (strategies 2 and 4) are shown in Fig. 2 for *M. didyma* and time period 1 (1991–2000). The results demonstrate that the exclusion of communes without species records did not lead to the hypothesized increase in model performance as measured by the adjusted D^2, max. Kappa and AUC values ($P = 1$). The results from *C. tullia* confirmed these findings, whereas for *M. teleius* partly significant increases were obtained (Fig. 2). However, these results were only obtained with the AUC with significant increases for strategy 2 vs. strategy 1 and strategy 4 vs. strategy 3 ($P < 0{\cdot}001$).

The comparisons between strategy 1 and strategy 3 and between strategy 2 and strategy 4 revealed the additional effects of not retaining presence data of auxiliary species in the pools of initial pseudo-absence data. The results for *M. didyma* and *M. teleius* in Fig. 2 were consistent with the findings achieved for *C. tullia*, with both strategy 3 and strategy 4 always significantly ($P < 0{\cdot}001$) outperforming strategy 1 and strategy 2, respectively.

HISTORICALLY DRIVEN PSEUDO-ABSENCE SELECTION

Considering presence data from longer time periods (T2 and T3), the resulting model performances of the four selection strategies were comparable to the results obtained using data from T1 (Table 4). Overall, strategies 1 and 3 outperformed strategies 2 and 4, respectively, except in the case of *M. teleius*, where model type S4–T2 significantly ($P < 0{\cdot}01$) outperformed S3–T2 using the AUC to quantify model performance.

The performances of *M. didyma* for the different historical data records are illustrated in Fig. 3. Using historical species records from the 100-year period (T3, 1901–2000) instead of records from only the last 10 years (T1, 1991–2000) significantly enhanced model performance ($P < 0{\cdot}001$) for every strategy used. Even the use of species data from T2 (1971–2000) instead of T1 increased model performances significantly ($P < 0{\cdot}001$). The same trend of increasing model performance was also achieved with *C. tullia* (Table 4). However, statistically non-significant ($P > 0{\cdot}05$) differences existed for the adjusted D^2 between S1–T3 and S1–T2 as well as for the AUC between S3–T2 and S3–T1. In the case of *M. teleius*, including longer historical data records significantly increased model performance of strategy 1 regarding the adjusted D^2 ($P < 0{\cdot}05$) and AUC ($P < 0{\cdot}01$) as well as of strategy 3 for all three performance measures ($P < 0{\cdot}001$). Nevertheless, in strategy 1 the Kappa values were significantly different ($P < 0{\cdot}05$) only between

Journal of Applied Ecology, **43**,
802–815

Table 4. Ranking of the different model types (MT) for the species according to the three model performance measures adjusted D^2, max. Kappa and AUC and the mean occurrence probabilities achieved for the evaluation data (new presence data 2001–05). Model performance decreases from top to bottom

Adjusted D^2					max. Kappa					AUC					Occurrence probabilities		
MT	P1	P2	Mean	SD	MT	P1	P2	Mean	SD	MT	P1	P2	Mean	SD	MT	Mean	SD
M. didyma (prevalence = 0·5)															Evaluation data (n = 33)		
S3–T3	***	***	0·644	0·051	S3–T3	***	***	0·772	0·037	S3–T3	***	***	0·963	0·010	S3–T3	0·66	0·34
S3–T2	***	***	0·610	0·051	S3–T2	***	***	0·748	0·038	S3–T2	***	***	0·956	0·011	S3–T2	0·64	0·34
S1–T3	NS	***	0·585	0·052	S4–T1.R	NS	*	0·729	0·036	S1–T3	NS	***	0·950	0·012	S1–T3	0·62	0·34
S4–T3	***	***	0·583	0·048	S4–T3	NS	*	0·728	0·037	S4–T3	***	***	0·950	0·011	S4–T1.R	0·62	0·35
S4–T2.R	NS	***	0·575	0·050	S4–T2.R	NS	***	0·726	0·040	S4–T2.R	NS	***	0·948	0·012	S4–T3	0·61	0·34
S4–T1.R	***	***	0·574	0·048	S1–T3	***	***	0·725	0·041	S4–T1.R	**	***	0·948	0·011	S4–T2.R	0·61	0·34
S3–T1	***	***	0·566	0·053	S3–T1	***	***	0·715	0·041	S3–T1	***	***	0·946	0·013	S3–T1	0·61	0·33
S1–T2	***	***	0·552	0·054	S1–T2	**	***	0·703	0·042	S1–T2	***	***	0·943	0·014	S1–T2	0·60	0·33
S4–T2	NS	***	0·537	0·049	S4–T2	*	***	0·698	0·040	S4–T2	NS	***	0·939	0·014	S1–T1	0·60	0·32
S1–T1	***	***	0·536	0·053	S1–T1	***	***	0·694	0·041	S1–T1	***	***	0·938	0·014	S4–T2	0·59	0·33
S2–T3	***	***	0·517	0·049	S2–T3	**	***	0·681	0·041	S2–T3	***	***	0·932	0·014	S2–T3	0·58	0·34
S2–T2.R	***	***	0·504	0·048	S2–T2.R	***	***	0·676	0·039	S2–T2.R	***	***	0·929	0·014	S2–T1.R	0·57	0·34
S2–T1.R	***	***	0·489	0·046	S2–T1.R	***	***	0·668	0·038	S2–T1.R	***	***	0·924	0·14	S2–T2.R	0·57	0·34
S2–T2	**	***	0·463	0·050	S4–T1	NS	***	0·646	0·040	S2–T2	NS	***	0·917	0·016	S4–T1	0·55	0·32
S4–T1	***		0·459	0·051	S2–T2	***		0·644	0·041	S4–T1	***		0·916	0·016	S1–T2	0·55	0·32
S2–T1			0·418	0·050	S2–T1			0·616	0·041	S2–T1			0·902	0·018	S2–T1	0·54	0·31
C. tullia (prevalence = 0·05)															Evaluation data (n = 7)		
S3–T3	***	***	0·671	0·149	S3–T3	***	***	0·658	0·078	S3–T3	***	***	0·968	0·026	S3–T3	0·054	0·05
S3–T2	***	***	0·635	0·128	S3–T2	***	***	0·644	0·081	S3–T2	NS	***	0·926	0·026	S4T1.R	0·051	0·03
S3–T1	**	***	0·596	0·119	S4–T3	*	***	0·629	0·092	S3–T1	***	***	0·961	0·025	S3–T2	0·049	0·04
S4–T3	NS	NS	0·580	0·124	S3–T1	***	***	0·621	0·080	S4–T3	NS	*	0·954	0·028	S4–T2.R	0·046	0·03
S4–T2.R	NS	NS	0·580	0·133	S4–T2	NS	NS	0·608	0·088	S1–T3	NS	*	0·952	0·032	S3–T1	0·045	0·04
S4–T1.R	NS	NS	0·576	0·135	S4–T2.R	NS	***	0·607	0·087	S4–T2.R	NS	NS	0·951	0·029	S1–T1	0·044	0·03
S1–T3	NS	NS	0·568	0·110	S1–T3	**	**	0·605	0·084	S4–T1.R	NS	NS	0·951	0·029	S2–T1	0·044	0·03
S4–T2	NS	***	0·567	0·122	S1–T2	NS	***	0·593	0·084	S1–T2	NS	***	0·950	0·031	S1–T2	0·043	0·03
S1–T2	***	***	0·563	0·108	S4–T1.R	***	***	0·593	0·086	S4–T2	*	***	0·948	0·030	S2–T1.R	0·043	0·02
S4–T1	**	***	0·521	0·093	S4–T1	***	***	0·577	0·087	S1–T1	***	***	0·946	0·030	S1–T3	0·043	0·03
S1–T1	**	***	0·508	0·093	S2–T3	NS	**	0·560	0·091	S4–T1	**	***	0·940	0·030	S4–T1	0·041	0·03
S2–T3	*	***	0·495	0·101	S1–T1	**	***	0·558	0·083	S2–T3	**	***	0·934	0·036	S4–T3	0·040	0·03
S2–T2.R	*	***	0·487	0·103	S2–T2.R	***	***	0·548	0·090	S2–T2.R	*	**	0·929	0·037	S2–T2	0·036	0·02
S2–T2	***	***	0·476	0·099	S2–T2	***	***	0·533	0·089	S2–T2	NS	***	0·927	0·036	S2–T3	0·035	0·02
S2–T1.R	***		0·462	0·090	S2–T1.R	***		0·515	0·079	S2–T1.R	***		0·925	0·039	S4–T2	0·035	0·02
S2–T1			0·395	0·071	S2–T1			0·452	0·078	S2–T1			0·915	0·033	S2–T2.R	0·034	0·02

Journal of Applied Ecology, **43**, 802–815

Table 4. *Continued*

Adjusted D^2					max. Kappa					AUC					Occurrence probabilities		
MT	P1	P2	Mean	SD	MT	P1	P2	Mean	SD	MT	P1	P2	Mean	SD	MT	Mean	SD
M. teleius (prevalence = 0·15)															Evaluation data (n = 21)		
S3–T3	**	***	0·436	0·072	S3–T3	***	***	0·618	0·058	S4–T1.R	***	***	0·905	0·026	S4–T1.R	0·53	0·29
S4–T1.R	*	***	0·428	0·075	S4–T1.R	NS	***	0·607	0·053	S4–T1	***	***	0·900	0·024	S4–T1	0·51	0·29
S3–T2	***	***	0·420	0·071	S3–T2	***	***	0·603	0·053	S3–T3	NS	NS	0·897	0·024	S3–T3	0·50	0·34
S4–T1	NS	NS	0·402	0·074	S3–T1	**	***	0·592	0·056	S4–T2	NS	**	0·896	0·025	S4–T2.R	0·49	0·31
S4–T2.R	NS	NS	0·399	0·077	S4–T3	***	***	0·585	0·055	S4–T2.R	*	**	0·895	0·026	S4–T2	0·49	0·31
S3–T1	NS	*	0·398	0·075	S4–T1	NS	***	0·574	0·053	S3–T2	NS	**	0·893	0·024	S3–T2	0·48	0·33
S4–T2	*	***	0·395	0·074	S4–T2.R	*	***	0·571	0·056	S2–T1.R	**	***	0·839	0·024	S4–T3	0·47	0·32
S4–T3	***	***	0·391	0·072	S1–T3	NS	NS	0·566	0·054	S4–T3	NS	*	0·889	0·025	S2–T1.R	0·47	0·26
S1–T3	*	***	0·379	0·067	S4–T2	NS	NS	0·563	0·055	S3–T1	NS	*	0·889	0·026	S3–T1	0·46	0·33
S1–T2	*	***	0·373	0·069	S1–T2	NS	***	0·563	0·054	S2–T1	NS	***	0·887	0·024	S2–T1	0·46	0·25
S2–T1.R	NS	***	0·367	0·066	S1–T1	***	***	0·560	0·055	S1–T3	*	**	0·886	0·025	S1–T3	0·46	0·31
S1–T1	***	***	0·363	0·070	S2–T3	NS	***	0·528	0·056	S1–T2	NS	NS	0·884	0·025	S1–T2	0·45	0·31
S2–T1	**	**	0·346	0·062	S2–T1.R	***	***	0·526	0·057	S2–T2	NS	NS	0·882	0·027	S2–T2	0·45	0·28
S2–T2	NS	***	0·339	0·067	S2–T2.R	NS	*	0·511	0·057	S2–T2.R	NS	**	0·881	0·027	S1–T2	0·44	0·30
S2–T2.R	NS		0·339	0·070	S2–T1	NS		0·509	0·057	S1–T1	*		0·881	0·026	S2–T2.R	0·44	0·28
S2–T3			0·338	0·064	S2–T2			0·506	0·056	S2–T3			0·878	0·026	S2–T3	0·43	0·29

P1, *P*-level comparing the model with the next in the sequence.
P2, *P*-level comparing the model with the second next in the sequence.
Significance levels: NS. > 0·05; * ≤ 0·05; ** ≤ 0·01; *** ≤ 0·001.

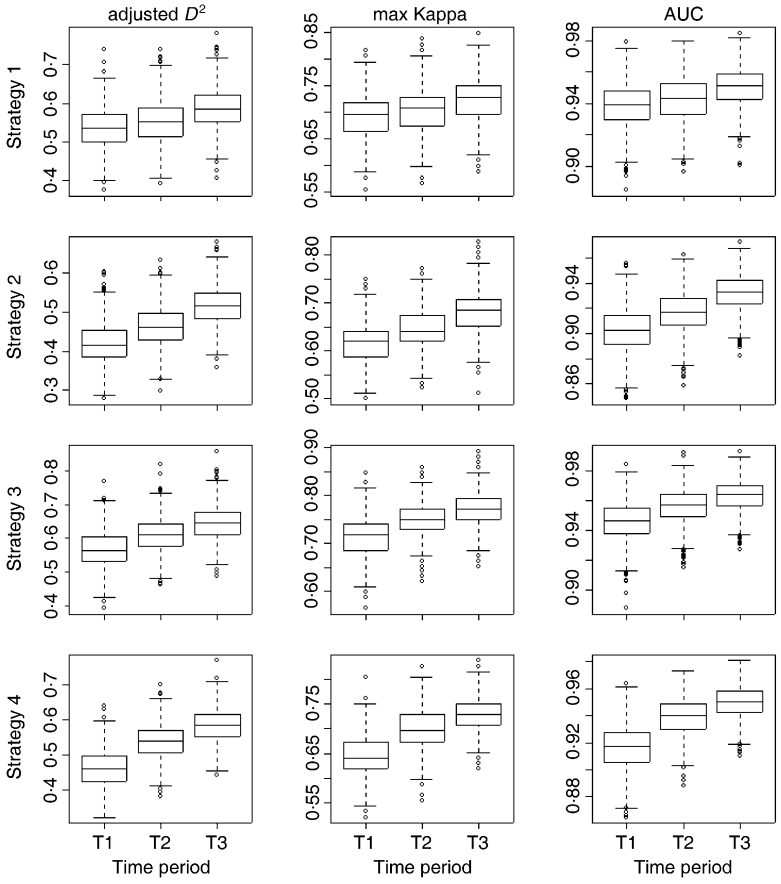

Fig. 3. Model performances for the three time periods T1 (1991–2000), T2 (1971–2000) and T3 (1901–2000) for *M. didyma.* Model performance is expressed by adjusted D^2, max. Kappa and AUC for the strategies S1, S2, S3 and S4.

T3 and T1. Strategy 2 and strategy 4 both showed a very similar pattern of performance. However, this pattern was often opposed to that of strategy 1 and strategy 3 showing a decrease in model performance when longer time periods were used. In fact, only the differences between T3 and T2 and between T3 and T1 for the Kappa value were statistically significant ($P < 0{\cdot}001$) for both strategies 2 and 4. All other pairwise comparisons were not significant ($P > 0{\cdot}85$), indicating a decrease in model performance and hence no advantage of taking into account historical data records.

Furthermore, the results for *M. didyma* revealed significant differences ($P < 0{\cdot}001$) between all pairs of restricted and non-restricted model types, indicating enhanced performance when former presence data did not remain in the pools of initial pseudo-absence data. For *C. tullia* the performance of the restricted strategies was significantly better in combination with T1 ($P < 0{\cdot}001$) as well as T2 ($P < 0{\cdot}05$), except for max. Kappa which did not differ significantly between S4–T2.R and S4–T2. No significance was found with *M. teleius* either for the adjusted D^2 and AUC for strategies 2 and 4 when data from T2 (1971–2000) were used. On the other hand, a significant increase in model performance ($P < 0{\cdot}001$) was achieved when data from T1 (1991–2000) were used. Significant ($P < 0{\cdot}05$) differences existed independent of the period the data originated from when restricted and non-restricted model types were compared by the max. Kappa coefficient.

SELECTION STRATEGIES VS. HISTORICAL SPECIES DATA

General model performance of the three species could be significantly enhanced by using longer time periods and refined strategies (Table 4). The adjusted D^2, measuring the proportion of explained deviance, ranged from 0·33 to 0·67, representing low to good model fits. Model accuracy measured with the AUC was high for all species, reaching values between 0·875 and 0·97. Even though max. Kappa was not the preferred measure for model accuracy because of its dependence on prevalence (Fielding & Bell 1997; McPherson, Jetz & Rogers 2004), we used it as a valuable performance measure to classify the model types because prevalence remained constant for species-specific models.

The ranking of the model types is compiled in Table 4 and shows the prominent model performances of strategy 3 and the benefits resulting from using species presence data from longer time periods. Strategy 2 was generally the least effective strategy and always clearly separated from strategy 3. Strategy 1 often revealed the next best model performances, followed or already mingled with the outcomes of strategy 4. Evaluating the benefits of the restricted pools of pseudo-absence data it could be recognized that for all three species the restricted pools of pseudo-absence data always led to better results than when only the non-restricted pools were used. However, in the case of *M. didyma* and *C. tullia* the restriction did not outperform model types using presence data from the longest time period T3. For every species the sequences of the model types across the three performance measures were compared using Spearman correlation coefficients (Table 5). An identical sequence ($R^2 = 1$) was only achieved between the adjusted D^2 and AUC for *M. didyma*, but the two measures also led to a similar sorting of the model types for *C. tullia*. Moreover, max. Kappa in combination with either the AUC or adjusted D^2 also provided good accordance. However, compared with *M. didyma* and *C. tullia*, *M. teleius* showed lower agreements between the performance measures. Furthermore, *M. teleius* also showed low agreements of the measures in combination with either *M. didyma* or *C. tullia* (Table 6), whereas the latter two showed sequences of model types more similar to each other. However, the max. Kappa value was very similar for every pairwise comparison, indicating that these sequences differed comparably.

The robustness of the resulting ranking of the model types on the basis of the three performance measures was confirmed using the occurrence probabilities for new species presence data recorded between 2001 and 2005 (Tables 4, 5 and 6). In particular, *M. didyma* showed

Table 5. Spearman rank correlation coefficients indicating the agreement among the rankings of the model types for each species. Rankings were based on the model performance measures (adjusted D^2, max. Kappa and AUC) and the occurrence probabilities (Prob.) predicted for new species presence data

	M. didyma			*C. tullia*			*M. teleius*		
	max. Kappa	AUC	Prob.	max. Kappa	AUC	Prob.	max. Kappa	AUC	Prob.
adj. D^2	0·97	1·00	0·98	0·97	0·99	0·56	0·92	0·88	0·89
max. Kappa		0·97	0·98		0·96	0·40		0·69	0·70
AUC			0·98			0·56			0·99

Table 6. Spearman rank correlation coefficients indicating the agreement among the rankings of the model types for pairs of species. Rankings were established for all performance measures (adjusted D^2, max. Kappa and AUC) and the occurrence probabilities predicted for new species presence (Prob.)

M. didyma vs. *C. tullia*		*C. tullia* vs. *M. teleius*		*M. didyma* vs. *M. teleius*	
adj. D^2	0·90	adj. D^2	0·81	adj. D^2	0·63
max. Kappa	0·88	max. Kappa	0·86	max. Kappa	0·86
AUC	0·94	AUC	0·40	AUC	0·31
Prob.	0·64	Prob.	0·50	Prob.	0·37

high accordance, whereas the lower values of the wetland species might have resulted from low prevalence.

Discussion

The two goals of this study were to compare different strategies of selecting pseudo-absence data from data sets that only contained species presence records, and to assess the role of historical species presence data in predicting species distribution. The strategies aimed at defining the most valid absence data by restricting the total number of communes present in the study area.

SELECTION STRATEGIES

The performance of a binomial generalized linear regression model increases with the degree with which presence data can be discriminated from absence data. In general, we would therefore expect valid absence data to discriminate better than pseudo-absence data. As the chance of selecting pseudo-absence data that results in high model performances is low when the whole data set is used, the performance of strategy 1 was expected to be the poorest. However, the results that were consistent for all species in the analyses showed that not considering communes lacking butterfly species presence data in strategies 2 and 4 resulted in significantly lower model performances compared with strategies 1 and 3, respectively. Therefore, in these cases only pseudo-absence data that discriminated presence from absence data less well were left in the pools. Hence the hypothesis that in communes where a lot of non-target species have been reported, the absence of a target species is expected to be more likely, could not be corroborated from our data. It seems, however, reasonable to think that large numbers of species records should indicate lower chances of having missed or not reported a presence of the target species. This suggests that, as communes where sufficient sampling effort was conducted were small in number, their role as well-discriminating pseudo-absence data might have been negligible in the analyses conducted here. Thus further analyses should be conducted to compare communes where only a few non-target species were recorded with communes where several non-target species were observed.

The hypothesized increase in model performance when communes with presence data of auxiliary species were no longer retained in the initial pools of pseudo-absence data (strategy 3 vs. strategy 1, strategy 4 vs. strategy 2) was confirmed by using new species presence data. Thus the discrimination power of the logistic regression models could be increased further by not considering communes in the pool of initial pseudo-absence data that shared similar environmental characteristics with the communes where the modelled species occurred.

Besides the strategies applied in this study, other strategies could possibly help to narrow down the number of pseudo-absence data. One alternative would be to let experts define sites where a certain species has never been observed and to use these sites as model absence data. Other selection strategies based on analytical approaches include the one proposed by Engler, Guisan & Rechsteiner (2004), where a prior model such as ENFA served to restrict pseudo-absence data to areas of low habitat suitability. However, the risk here is that a bias in the stratifying ENFA model becomes amplified in the binomial presence–absence GLM. Furthermore, pseudo-absence data could be preferably sampled from sites with environmental characteristics not similar to those where the original presence data were found (Zaniewski, Lehmann & Overton 2002). Some authors have pointed to possible negative effects as a result of the inclusion of absence data from beyond the species' known distribution (Austin & Meyers 1996; Thuiller *et al*. 2004). The strategies in this study did not account for such absence data because the range of species' distribution was rather wide compared

with the extent of the study area and related environmental range. However, the results of the wetlands species showed that the models could still be improved concerning their predictive success, which might result from the low prevalence of these species.

HISTORICAL SPECIES DATA

The second goal was to analyse the benefits of using historical species records comprising different time periods for predicting species distribution. The results revealed that the consideration of longer time periods generally greatly enhanced model performance. Significant enhancements were already achieved when only former presence communes of the target species itself did not remain in the pools of pseudo-absence data, as shown in strategies 1 and 2. Model performance was also significantly increased in strategies 3 and 4 where additionally the former presence of auxiliary species was not considered. An increase in model performance resulted when more communes for which environmental characteristics differed from communes with species presence remained in the initial pools of pseudo-absence data. Hence it can be concluded that the present-day environmental characteristics of communes where recent presence was recorded are very similar to communes where presence was mentioned in the past. Consequently, it is possible that a species presence will be reconfirmed in communes where it was once observed. This reveals the problem of modelling species' distributions by using unsystematically recorded presence-only data that just contains records from the most recent time period, for example the last 10 years, and where 'ghosts' of past presence can take on great importance. As not all possible presence data were recorded in this period, sample bias, a problem inherent to most presence-only NHC data sets (Hirzel *et al*. 2002; Graham *et al*. 2004), can be expected. Yet the strategies presented here make active use of former presence data by not considering them as pseudo-absence data in the GLM.

For *M. teleius*, increases in model performance in strategies 2 and 4 were not always achieved by using data originating from longer time periods. As the number of pseudo-absence data in T2 only differed slightly between *C. tullia* and *M. teleius* and was actually identical in T4, we reject the possibility that the number of pseudo-absence data in the initial pools caused the different results. Instead we attribute the encountered deviance to the quality of the presence data of *M. teleius* used in the models. However, further analyses have to be conducted to analyse the influence of different presence data on model performance.

COMPARISON BETWEEN STRATEGIES AND HISTORICAL RECORDS

The ranking of the model types according to the three performance measures revealed that changing the strategy often contributed more to enhance model performance than considering historical species records. This applied most of all to *C. tullia* and *M. didyma*. These two species also showed very similar rankings of the model types, independent of the performance measure used. Moreover, the order of the model types was very similar between these two species, whereas the comparisons with *M. teleius* revealed lower accordance. We interpret the lower correlation values obtained for *M. teleius* with its general lower model performance. The highest values of adjusted D^2 and AUC achieved for *M. teleius* approximated the lowest values achieved for *C. tullia* and *M. didyma*.

MANAGEMENT IMPLICATIONS

Species distribution models have become very important management tools in nature conservation planning (Rushton, Ormerod & Kerby 2004; Guisan & Thuiller 2005; Whittaker *et al*. 2005). Such models can help planning authorities detect new species occurrences (Engler, Guisan & Rechsteiner 2004; Guisan & Thuiller 2005) and define priority areas for species protection or reintroduction (Pearce & Lindenmayer 1998; Hirzel *et al*. 2002). Up to the present, the value of historic species presence data originating from abundant NHC has received little attention in connection with modelling purposes. However, these data have been shown to be useful for fitting predictive species distribution models (Graham *et al*. 2004). The results of this study show that models using NHC data to select pseudo-absence data can be improved compared with models that rely on purely random selections of pseudo-absence data. Therefore, we suggest using (i) recent presence data of auxiliary species and (ii) historical presence data of the modelled and auxiliary species to restrict the pools of pseudo-absence data used to fit presence–absence models. These findings need further testing with other data in other areas. None the less, our results confirm that saving historical species distribution data from NHC should become a task of high priority for future research in conservation biology and applied ecology.

Acknowledgements

This study was supported by a grant from the Swiss National Science Foundation (application no. 4048–064460) in the program 'Landscapes and Habitats of the Alps' (NRP 48). We are thankful to Yves Gonseth, Simon Capt and Anthony Lehman from the Swiss Centre for Faunal Cartography (CSCF) in Neuchâtel for providing the species data as well as Daniel Bohnenblust, Marianne Gerber and Frédéric In-Albon from the Swiss Federal Statistical Office in Neuchâtel for recent census data. We thank Janine Bolliger for constructive discussion on previous versions of the manuscript. We gratefully acknowledge Dr Rob Freckleton and two anonymous referees for their valuable comments that significantly improved the manuscript.

References

Austin, M.P. & Meyers, J.A. (1996) Current approaches to modelling the environmental niche of eucalypts: implication for management of forest biodiversity. *Forest Ecological Management*, **85**, 95–106.

Busby, J.R. (1991) BIOCLIM: a bioclimate analysis and prediction system. *Nature Conservation: Cost Effective Biological Surveys and Data Analysis* (eds C.R. Margules & M.P. Austin), pp. 64–68. CSIRO, Melbourne, Australia.

Cohen, J. (1960) A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, **20**, 37–46.

Dirnböck, T., Dullinger, S. & Grabherr, G. (2003) A regional impact assessment of climate and land-use change on alpine vegetation. *Journal of Biogeography*, **30**, 401–417.

Ebert, G. & Rennwald, E. (1993) *Die Schmetterlinge Baden-Württembergs*. Ulmer, Stuttgart, Germany.

Engler, R., Guisan, A. & Rechsteiner, L. (2004) An improved approach for predicting the distribution of rare and endangered species from occurrence and pseudo-absence data. *Journal of Applied Ecology*, **41**, 263–274.

Fielding, A.H. & Bell, J.F. (1997) A review of methods for the assessment of prediction errors in conservation presence/absence models. *Environmental Conservation*, **24**, 38–49.

Gonseth, Y. (1987) *Verbreitungsatlas der Tagfalter der Schweiz (Lepidoptera Rhopalocera)*. Centre Suisse de Carthographie de la Faune, Neuchâtel, Switzerland.

Gonseth, Y. (1994) Rote Liste der gefährdeten Tagfalter der Schweiz. *Rote Listen der Gefährdeten Tierarten in der Schweiz* (ed. P. Duelli), pp. 48–51. EMDZ, Bern, Switzerland.

Graham, C.H., Ferrier, S., Huettman, F., Moritz, C. & Peterson, A.T. (2004) New developments in museum-based informatics and applications in biodiversity analysis. *Trends in Ecology and Evolution*, **19**, 497–503.

Guisan, A. & Hofer, U. (2003) Predicting reptile distributions at the mesoscale: relation to climate and topography. *Journal of Biogeography*, **30**, 1233–1243.

Guisan, A. & Theurillat, J.P. (2000) Equilibrium modeling of alpine plant distribution: how far can we go? *Phytocoenologia*, **30**, 353–384.

Guisan, A. & Thuiller, W. (2005) Predicting species distribution: offering more than simple habitat models? *Ecology Letters*, **8**, 993–1009.

Guisan, A. & Zimmermann, N.E. (2000) Predictive habitat distribution models in ecology. *Ecological Modelling*, **135**, 147–186.

Hirzel, A.H., Hausser, J., Chessel, D. & Perrin, N. (2002) Ecological-niche factor analysis: how to compute habitat-suitability maps without absence data? *Ecology*, **83**, 2027–2036.

Hirzel, A.H., Helfer, V. & Metral, F. (2001) Assessing habitat-suitability models with a virtual species. *Ecological Modelling*, **145**, 111–121.

Houlder, D.J., Hutchinson, M.F., Nix, H.A. & McMahon, J.P. (2000) *ANUCLIM User Guide, Version 5·1.* Centre for Resource and Environmental Studies, Australian National University, Canberra, Australia. http://cres.anu.edu.au/outputs/anuclim/doc/Contents.html (accessed 15 May 2006).

Hutchinson, G.E. (1957) Concluding remarks. *Cold Spring Harbour Symposium. Quantitative Biology*, **2**, 415–427.

Jaberg, C. & Guisan, A. (2001) Modelling the distribution of bats in relation to landscape structure in a temperate mountain environment. *Journal of Applied Ecology*, **38**, 1169–1181.

Kéry, M. (2002) Inferring the absence of a species: a case study of snakes. *Journal of Wildlife Management*, **66**, 330–338.

McArdle, B.H. (1990) When are rare species not there? *Oikos*, **57**, 276–277.

McCullagh, P. & Nelder, J.A. (1989) *Generalized Linear Models*, 2nd edn. Chapman & Hall, London, UK.

MacKenzie, D.I., Nichols, J.D., Hines, J.E., Knutson, M.G. & Franklin, A.B. (2003) Estimating site occupancy, colonization, and local extinction when a species is detected imperfectly. *Ecology*, **84**, 2200–2207.

McPherson, J.M., Jetz, W. & Rogers, D.J. (2004) The effects of species' range sizes on the accuracy of distribution models: ecological phenomenon or statistical artefact? *Journal of Applied Ecology*, **41**, 811–823.

Maggini, R., Lehmann, A., Zimmermann, N.E. & Guisan, A. (2006) Improving generalized regression analysis for the spatial prediction of forest communities. *Journal of Biogeography*, DOI: 10.1111/j.1365-2699.2006.01465.x.

Manly, B.F.J. (1997) *Randomization, Bootstrap and Monte Carlo Methods in Biology*, 2nd edn. Chapman & Hall, London, UK.

Osborne, P.E., Alonso, J.C. & Bryant, R.G. (2001) Modelling landscape-scale habitat use using GIS and remote sensing: a case study with great bustards. *Journal of Applied Ecology*, **38**, 458–471.

Pearce, J. & Lindenmayer, D. (1998) Bioclimatic analysis to enhance reintroduction biology of the endangered helmeted honeyeater (*Lichenostomus melanops cassidix*) in southeastern Australia. *Restoration Ecology*, **6**, 238–243.

Pearson, S.M., Turner, M.G. & Drake, J.B. (1999) Landscape change and habitat availability in the Southern Appalachian Highlands and Olympic Peninsula. *Ecological Applications*, **9**, 1288–1304.

Peterson, A.T., Egbert, S.L., Sanchez-Cordero, V. & Price, K.P. (2000) Geographic analysis of conservation priority: endemic birds and mammals in Veracruz, Mexico. *Biological Conservation*, **93**, 85–94.

Robertson, M.P., Caithness, N. & Villet, M.H. (2001) A PCA-based modelling technique for predicting environmental suitability for organisms from presence records. *Diversity and Distributions*, **7**, 15–27.

Rushton, S.P., Ormerod, S.J. & Kerby, G. (2004) New paradigms for modelling species' distributions? *Journal of Applied Ecology*, **41**, 193–200.

Schuler, M. (1997) *Die Raumgliederungen der Schweiz*. Bundesamt für Statistik, Bern, Switzerland.

Schuler, M., Ullmann, D. & Haug, W. (2002) *Bevölkerungsentwicklung der Gemeinden 1850–2000. Eidgenössische Volkszählung 2000*. Bundesamt für Statistik, Neuchâtel, Switzerland.

Swisstopo (2005) *Digital Elevation Model 25 Level 2*. Federal Office of Topography, Wabern, Switzerland.

Thuiller, W., Brotons, L., Araujo, M.B. & Lavorel, S. (2004) Effects of restricting environmental range of data to project current and future species distributions. *Ecography*, **27**, 165–172.

Walter, T. & Schneider, K. (2003) Eco-fauna-database: a tool for both selecting indicator species for land use and estimating impacts of land use on animal species. *Agriculture and Biodiversity. Developing Indicators for Policy Analysis* (eds Organisation for economic co-operation and development (OECD)), pp. 152–155. OECD, Paris, France.

Whittaker, R.J., Araujo, M.B., Paul, J., Ladle, R.J., Watson, J.E.M. & Willis, K.J. (2005) Conservation biogeography: assessment and prospect. *Diversity and Distributions*, **11**, 3–23.

Zaniewski, A.E., Lehmann, A. & Overton, J.M.C. (2002) Predicting species spatial distributions using presence-only data: a case study of native New Zealand ferns. *Ecological Modelling*, **157**, 261–280.

Zimmermann, N.E. & Kienast, F. (1999) Predictive mapping of alpine grasslands in Switzerland: species versus community approach. *Journal of Vegetation Science*, **10**, 469–482.

Received 12 August 2005; final copy received 18 March 2006
Editor: Rob Freckleton